D346–D354 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 13 November 2009*
doi:10.1093/nar/gkp848

The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata

Konstantinos Liolios[1], I-Min A. Chen[2], Konstantinos Mavromatis[1], Nektarios Tavernarakis[3], Philip Hugenholtz[4], Victor M. Markowitz[2] and Nikos C. Kyrpides[1,*]

[1]Genome Biology Program, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, [2]Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, [3]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology, Heraklion, Crete, Greece and [4]Microbial Ecology Program, DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, USA

Received September 19, 2009; Accepted September 22, 2009

ABSTRACT

The Genomes On Line Database (GOLD) is a comprehensive resource for centralized monitoring of genome and metagenome projects worldwide. Both complete and ongoing projects, along with their associated metadata, can be accessed in GOLD through precomputed tables and a search page. As of September 2009, GOLD contains information for more than 5800 sequencing projects, of which 1100 have been completed and their sequence data deposited in a public repository. GOLD continues to expand, moving toward the goal of providing the most comprehensive repository of metadata information related to the projects and their organisms/environments in accordance with the Minimum Information about a (Meta)Genome Sequence (MIGS/MIMS) specification. GOLD is available at: http://www.genomesonline.org and has a mirror site at the Institute of Molecular Biology and Biotechnology, Crete, Greece, at: http://gold.imbb.forth.gr/

HISTORY AND GROWTH

The Genomes OnLine Database (GOLD) provides a centralized resource for the continuous monitoring of genome and metagenome sequencing projects worldwide, uniquely integrated with their associated metadata. Since its founding in 1997 (1–4), GOLD has grown dramatically, now hosting information regarding over 5800 sequencing projects (Figure 1A).

The number of registered sequencing projects has doubled since the publication of the previous report two years ago (4). As of September 2009, 5843 projects have been recorded, versus 2905 as of September 2007 and 1575 as of September 2005 (3, 4). This rapid growth has been fueled by decreasing sequencing costs combined with technological advances, and was significantly augmented by the launching and successful execution of several large-scale microbial genome sequencing initiatives, e.g. the Human Microbiome Project (http://www.hmpdacc.org/) and the Genomic Encyclopedia of Bacteria and Archaea (http://www.jgi.doe.gov/programs/GEBA/). During this period, GOLD has also expanded its scope beyond standard genomic and metagenomic projects to now encompass data from the growing number of resequencing, transcriptome and metatranscriptome projects.

In parallel with this doubling in the number of genome projects has come an increase in the number of captured metadata fields from 56 in 2007 (4) to 135 today. This is an area of active development; thus, we anticipate further increases as more metadata types are described and captured in published studies. Some of the new metadata types are described below.

Among the most important developments of the database during the last 2 years are those coupled to the growth of the metadata. These include the implementation of GOLD-specific Controlled Vocabularies (CVs) for the representation of the associated data, as well as coordination with the Genomics Standards Consortium (GSC) (http://gensc.org/) and compliance with its recommendations for the Minimum Information about a (Meta)Genome Sequence (MIGS/MIMS) (5).

As the rate of launching new projects accelerates, the task of monitoring and recording their data along with their metadata grows ever more difficult. Therefore, the sequencing centers and the community at large are strongly encouraged to register their own sequencing projects in GOLD to ensure complete and accurate project tracking.

*To whom correspondence should be addressed. Tel: +1 925 296 5718; Fax: +1 925 296 5850; Email: nckyrpides@lbl.gov

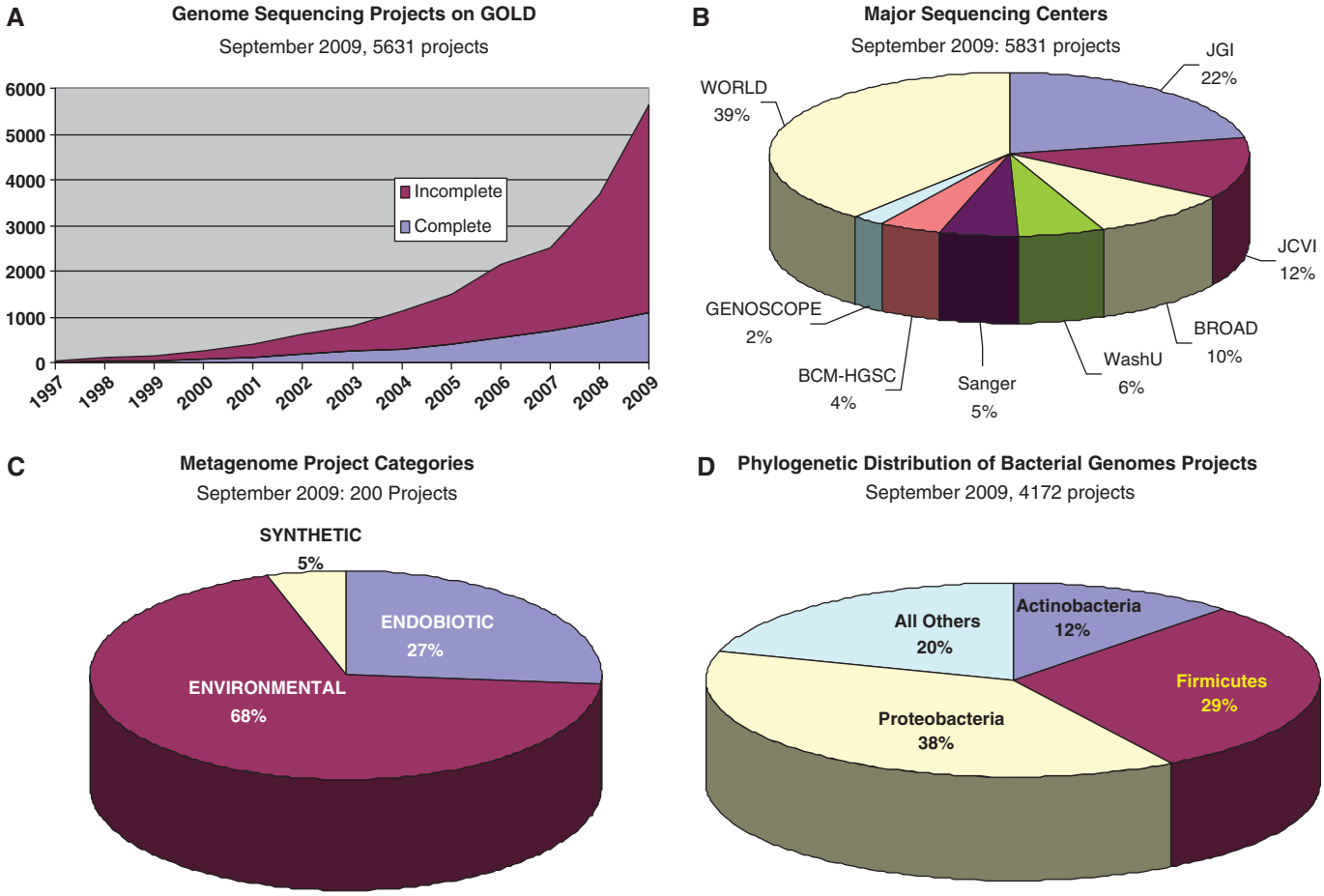


Figure 1. Statistical information available from GOLD. (**A**) Evolution of the complete and ongoing genome projects monitored in GOLD from December 1997 through September 2009. (**B**) Distribution of the 5831 genome projects across the major sequencing centers. Abbreviations: JGI, Joint Genome Institute; JCVI, J. Craig Venter Institute; Broad, Broad Institute; WashU, Washington University; Sanger, the Wellcome Trust Sanger Institute; BCM-HGSC, Baylor College of Medicine Human Genome Sequencing Center; WORLD, all other sequencing centers. (**C**) Distribution of the 200 current metagenome projects across the three major metagenome classification categories. (**D**) Phylogenetic distribution of the 4172 bacterial genome projects as of September 2009.

CURRENT STATUS OF GOLD

Published complete genomes

The year 2009 represents a landmark in the history of genome sequencing projects: the completed sequencing of the first 1000 genomes. As of September 2009, GOLD documents 1100 completed genome projects, a 1.7-fold increase from 2 years ago (4). These comprise 914 bacterial, 68 archaeal and 118 eukaryotic genomes. Thus, the completely sequenced archaeal and bacterial genomes currently total 982, leading one to confidently predict that the community will celebrate yet another 1000 genome milestone before the end of the year.

For all of these projects, the complete genome sequence is 'published' by being deposited in one of the public archival databases such as GenBank (6), EMBL (7) and DDBJ (8). However, a rapidly increasing proportion of the projects do not have an associated publication in the literature. That fraction currently stands at 37% (408 of 1100). This shift is partly attributable to the more frequent release of sequence data to the community prior to publication in compliance with the rapid pre-publication data release policies and recommendations (9). Another factor is the increase in larger-scale efforts that involve the parallel sequencing of several hundred organisms (e.g. the HMP and GEBA). Here, preparation of the typical detailed publication describing the genome of every single organism would be virtually impossible (4,10). This situation calls for a new mechanism that can provide a GSC-compliant citable record for every completed genome project and its metadata. To that end, an open access scientific journal, *Standards in Genomic Sciences* (SIGS), (http://standardsingenomics.org/) has recently been launched (11), its goal being to catalog and maintain the data from completed genome projects in an orderly and standardized manner (10).

In addition to publication of each complete genome sequence, GSC also strongly recommends that the source organism be available from a culture collection center. It is unfortunate that after so many years and so many genome sequences, the widely accepted policies for publication of genome sequencing projects require the submission to a public repository of only the sequence data, not also the biological material itself. As a result,

Table 1. Project type distribution[a]

Archaea: 179	Genome: 169	Transcriptome: 0	Resequencing: 0	Uncultured: 9
Bacteria: 4184	Genome: 4097	Transcriptome: 4	Resequencing: 35	Uncultured: 14
Eukarya: 1280	Genome: 804	EST/Transcriptome: 344	Resequencing: 45	Uncultured: 1

[a]Available at: http://www.genomesonline.org/project_type_distribution.cgi.

Table 2. Project status distribution[a]

Archaea: 179	Complete: 74	Draft: 16	In Progress: 17	Awaiting DNA: 7	Targeted: 1
Bacteria: 4184	Complete: 1151	Draft: 950	In Progress: 414	Awaiting DNA: 142	Targeted: 517
Eukarya: 1280	Complete: 159	Draft: 178	In progress: 46	Awaiting DNA: 73	Targeted: 9

[a]Available at: http://genomesonline.org/sequencing_status_distribution.cgi.

from the current list of 982 completed archaeal and bacterial genomes, only 518 (53%) appear to be available from a culture collection center (12), and only half of those genomes (27% of the total) represent a type strain of the sequenced species.

Ongoing genome projects

In addition to the 1095 completed projects, there are currently 4543 ongoing sequencing projects, of which 3271 are bacterial, 110 archaeal and 1162 are eukaryotic. This total is more than double the 2158 reported 2 years ago. Until recently, the projects monitored for GOLD were predominantly 'Genome' and 'EST' sequencing projects, supplemented by a small number of 'Genome-Surveys' and 'Genome-Regions' (the latter representing some eukaryotic projects focused on specific genomic regions). The increasing number of 'Resequencing' and 'Transcriptome' projects prompted the addition of these two new project types during the past year (Table 1).

The current Sequencing Status distribution tallied by domain is shown in Table 2. The Sequencing Status designations and current tallies are as follows:

- *Complete*: DNA sequencing has been completed; 288 projects in addition to the 1100 already published.
- *Draft*: a draft sequence has been deposited in a public repository; 1164 projects.
- *In progress*: the DNA sequence has been received by the sequencing center but there is not yet public data release; 442 projects.
- *Awaiting DNA*: an organism selection has been made, but the DNA has not yet arrived at the DNA sequencing center; 236 projects.
- *Targeted*: a project has been identified but further work has not yet begun; 527 projects.

The distributions of all projects by Project Type and by Sequencing Status are now dynamically tracked with every GOLD update and can be viewed online through the main page at: http://www.genomesonline.org/gold.cgi.

Metagenome projects

The past 2 years have seen a growing number of metagenomic projects added to GOLD, and the expectation is that this trend will continue, reinforced by further advances in the sequencing technology. The database currently reports 200 distinct metagenomic projects, embracing 453 samples.

During curation, careful attention is paid to ensure that project names follow the standardized schema previously described (4). All the metagenome projects are classified under three major categories: environmental (137 projects), endobiotic or host-associated (53 projects) and synthetic (10 samples) (Figure 1C). A project classification schema is also under development and will soon be released from the database. A prototype of this classification has already been adopted by the Integrated Microbial Genomes with Microbiome Samples (IMG/M) database (13) and is available for browsing online (http://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonList&page=taxonListPhylo&domain=*Microbiome&genome_type=metagenome). A hierarchical classification scheme with all the metagenome projects captured in GOLD will soon be available from the database.

Metadata

The genome/metagenome associated metadata have also undergone significant expansion in GOLD during the last 2 years. The number of metadata categories has increased from two in the previous release to six in GOLD v.3: (i) organism information; (ii) project information; (iii) sequencing information; (iv) environmental metadata; (v) host metadata; and (vi) organism metadata. Likewise, the number of metadata fields assigned to those categories has grown from 56 to 135.

The current status of the different fields and the number of projects with associated data for each of the corresponding fields is shown in Table 3. Some of the metadata fields are populated for all or most of the projects, while other fields (particularly newer ones) are yet to be curated for the majority of the projects. Although the number of metadata fields is expected to continue to grow, the current list has already been put to use in microbial comparative analysis systems such as the Integrated Microbial Genomes IMG (14) and IMG/M (13).

Table 3. Metadata categories and fields

1. Organism information	Type	No. of projects	2. Project information	Type	No. of projects
1. GOLD display name	FT	5843	1. GOLD project ID	ID	5843
2. NCBI project Name	FT	3408	2. GCAT ID	ID	5843
3. Common name	FT	364	3. NCBI project ID	ID	3600
4. Domain	CV	5843	4. IMG ID	ID	1664
5. Phylum	CV	5665	5. Cross reference ID	ID	204
6. Class	CV	5379	6. Greengenes ID	ID	1994
7. Order	CV	5608	7. 16S ID	ID	17
8. Family	CV	5396	8. NCBI archive ID	ID	15
9. Genus	CV	5570	9. Short read archive ID	ID	117
10. Species	CV	3856	10. Project type	CV	5843
11. Strain	FT	4748	11. Project status	CV	5843
12. Serovar	FT	384	12. Availability	CV	5843
13. NCBI taxon ID	ID	5699	13. Contact name	FT	4210
14. Culture collection ID	FT	1711	14. Contact email	FT	3480
15. Type strain	CV	1970	15. Contact link	URL	1034
16. Biosafety level	ID	260	16. Funding program	CV	1612
17. Organism comments	FT	11	17. Proteomics data	FT	2
			18. Proteomics Link	URL	2
3. Sequencing information			19. Transcriptomics Data	FT	14
1. Sequencing Status	CV	3870	20. Transcriptomics Link	URL	5
2. Sequencing quality	CV	321	21. Locus Tag	FT	1286
3. Seq status link	URL	800	22. GC percent	FT	2380
4. Library method	FT	134	23. Chromosome count	ID	1259
5. Number of reads	FT	173	24. Plasmid count	ID	1223
6. Vector	FT	65	25. Completion date	ID	1155
7. Assembly method	FT	368	26. Publication	CV	1154
8. Sequencing depth	FT	1277	27. Project description	FT	205
9. Gene calling method	FT	263	28. Project relevance	CV	10 396
10. Contig count	FT	583	29. Funding center	CV	4450
11. Estimated size	FT	2780	39. Sequence data	ID	2794
12. Gene count	FT	1993	31. Database	CV	5101
13. Sequencing country	CV	5802			
			6. Organism metadata		
4. Environmental metadata			1. Oxygen requirement	CV	3797
1. Isolation site	FT	3188	2. Cell shape	CV	3710
2. Source of isolate	FT	609	3. Motility	CV	3435
3. Method of isolation	FT	141	4. Sporulation	CV	2610
4. Isolation comments	FT	134	5. Temperature Range	CV	4422
5. Collection date	FT	426	6. Temperature optimum	ID	1319
6. Isolation country	CV	1345	7. Salinity	CV	131
7. Isolation Pubmed ID	ID	104	8. pH	ID	180
8. Geographic iocation	FT	2138	9. Cell diameter	FT	68
9. Latitude	FT	769	10. Cell length	FT	56
10. Longitude	FT	768	11. Color	CV	44
11. Altitude	FT	16	12. Gram staining	CV	4229
12. Depth	FT	193	13. Biotic relationships	CV	4244
			14. Symbiotic physical interaction	CV	135
5. Host metadata			15. Symbiotic relationship	CV	182
1. Host name	FT	2029	16. Symbiont name	FT	156
2. Host gender	FT	219	17. Cell arrangement	CV	1897
3. Host race	FT	3	18. Diseases	CV	5303
4. Host age	FT	143	19. Habitat	CV	7214
5. Host health	FT	363	20. Metabolism	CV	21
6. Host medication	FT	2	21. Phenotypes	CV	3345
7. Primary body sample site	CV	1643	22. Energy source	CV	1439
8. Body sample subsite	CV	533			
9. Body product	CV	412			
10. Additional body sample site	CV	18			

Abbreviations for field types: ID, identity number; FT, free text; CV, control vocabulary; URL, uniform resource locator.

Particularly important developments currently underway involve the integration and mapping of several of the available metadata fields in GOLD to well-developed publicly available metadata ontologies and control vocabularies such as 'Habitat-Lite' (15) and others.

NEW DEVELOPMENTS

New user interface implementing new technologies

The burgeoning array of new types of data recorded in GOLD necessitated a major revamping of the graphical user interface. The GOLD tables have been visually

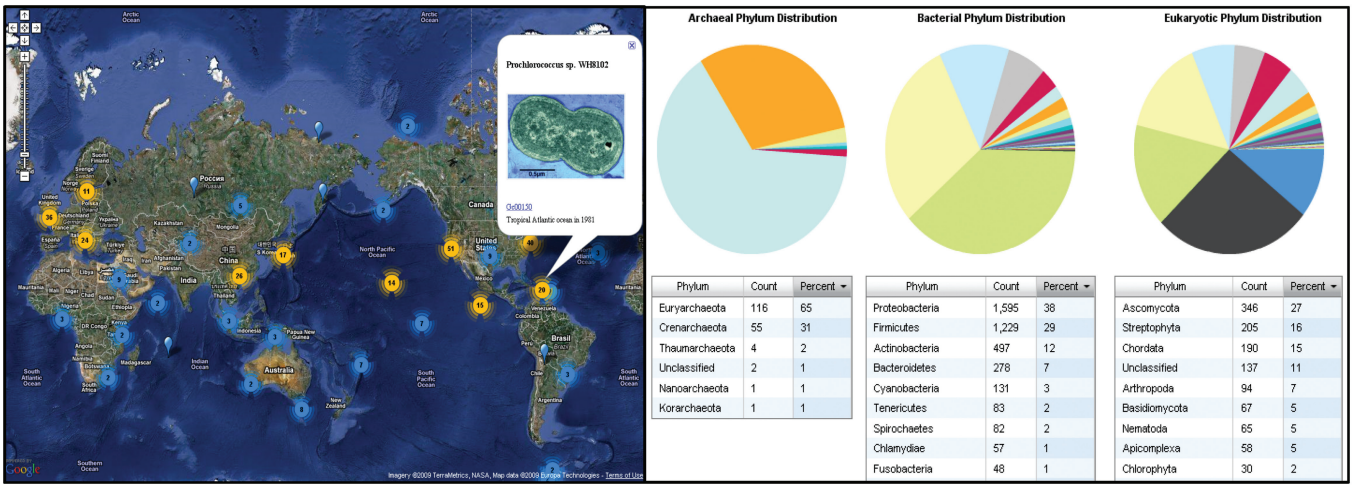


Figure 2. Graphical displays in GOLD. (Left) Geographical display of the collection location for organisms and environmental samples. Click on a project to view the detailed information window showing the name of the project, an image (if available), a GOLD CARD link, and a short description identifying the location. (Right) Phylogenetic distribution of archaeal, bacterial and eukaryotic projects with accompanying data tables.

enhanced using advanced graphical technologies such as EXT JS JavaScript library for the grids, the Yahoo User Interface Library for the pie charts and data tables, the Google Maps API for geographical location display, Google MarkerClusterer for improved visual display of multiple map locations, and the JavaScript Object Notation (JSON) data format for rapid data loading.

On the main page (http://www.genomesonline.org/gold.cgi), three links have been added to connect to new pages displaying the current distribution of projects by type, sequencing status and phylogeny (Figure 2, right). On each of these new pages, the same technologies are used to convey key breakdown data in a visually intuitive manner. Below the links, the Google map is displayed showing all projects individually or in clusters (Figure 2, left). Clicking on a project displays information about the collection location, an image (if available), and a link to the project's GOLD CARD page.

The same entry page provides access to the enhanced tables for the five major GOLD project categories (published complete genomes, archaeal ongoing genomes, bacterial ongoing genomes, eukaryotic ongoing genomes and metagenomes). Each table displays information for 12 primary metadata fields for each project. By default, projects are sorted by GOLDSTAMP ID, sequential numbers assigned in sequence as projects are entered in GOLD. To sort by the data in any other column, click the column header. To display advanced options, mouse over the column header and click to open the dropdown list. These options enable you to sort in ascending or descending order, to show/hide different columns and to filter the projects displayed based on data in that column.

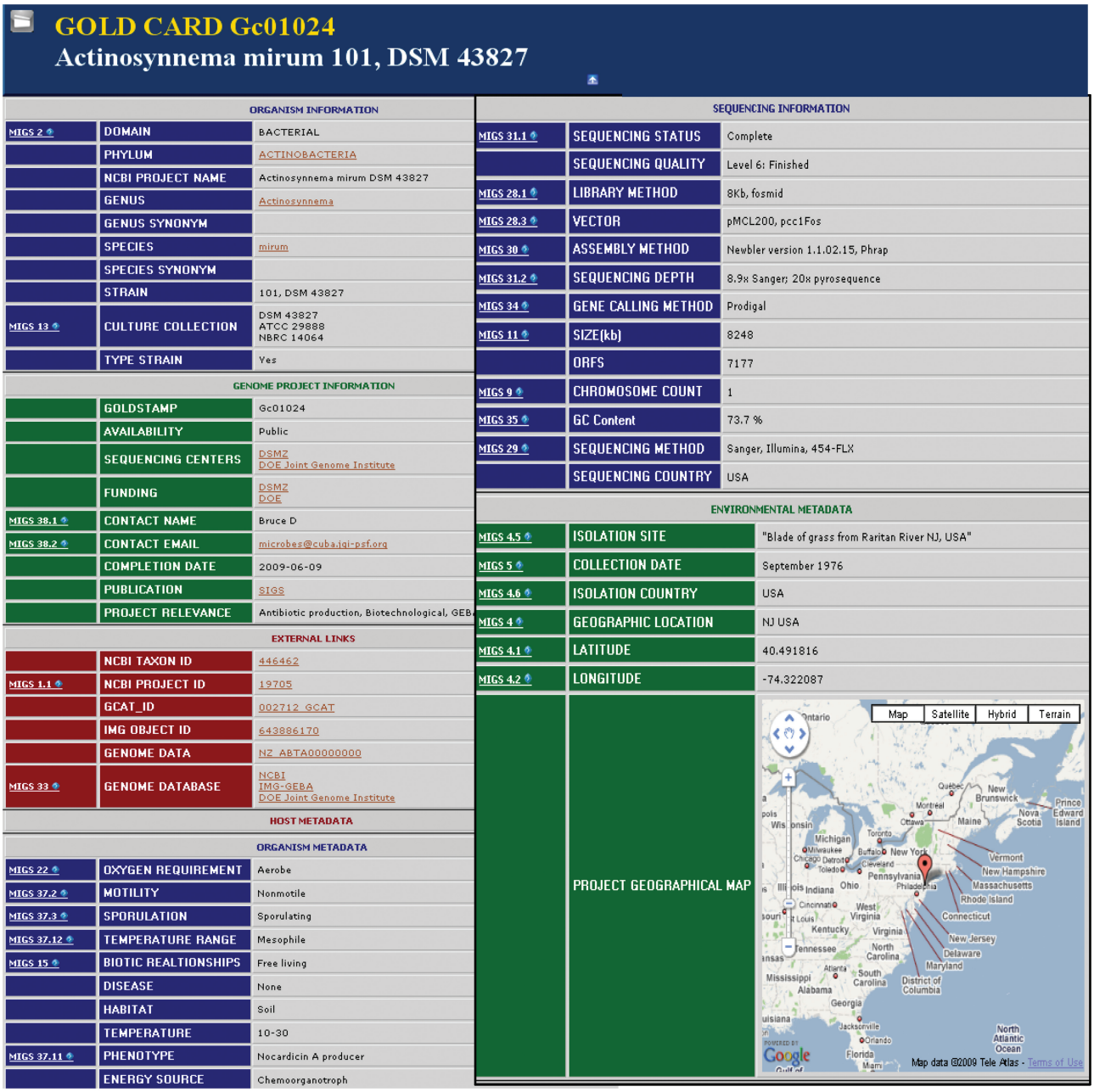
The Search GOLD page has been completely rewritten. There are currently four tab pages, each corresponding to a different search mode and each offering new capabilities for more effective searching. The first tab, the basic search, provides commonly used Boolean queries for the most frequently searched fields in three main data categories. The Advanced Search tab offers a more extended list of search criteria from eight major data categories. The Metadata Search tab can be used to query the database metadata and view the results in tables and graphical displays of statistics and rankings. A fourth tab that is currently under development, Custom (SQL) Search, will enable users to construct and execute their own SQL queries. The aforementioned interface technologies are also employed here to provide an enhanced visual display of the search results and enable further manipulations. The user can export the search results to a Microsoft® Excel file or redirect them to the metadata analysis page. At that page, charts and statistics can be derived from the breakdown of the search results based on more than 40 metadata fields.

Finally, the GOLD CARD page has also been extensively redesigned, making for more intuitive navigation (Figure 3). Genome project data are now organized into seven major categories for easier access. Google map location and images of the organism(s) are provided when available. Empty data rows can be hidden by clicking the arrow located at the upper right corner of the card. The GOLD CARD page complies with the GSC standards (5) and provides IDs and links for all the compliant data fields. The list of metadata fields provided by GOLD, now more than 100, includes those currently part of the MIGS specifications plus many more that are now candidates for inclusion in the MIGS list.

The prefix in the GOLDSTAMP identifier assigned to each project encodes additional project information: Gc, GOLD complete; Gi, GOLD incomplete; Gm, GOLD metagenome; Ge, GOLD EST; Gr, GOLD resequencing; and Gt, GOLD transcriptome.

Metadata collection and management system

The number of genome projects initiated is increasing exponentially, bringing with it an exponential increase in the task of curating the GOLD data. To help cope with



GOLD CARD Gc01024
Actinosynnema mirum 101, DSM 43827

ORGANISM INFORMATION		
MIGS 2	DOMAIN	BACTERIAL
	PHYLUM	ACTINOBACTERIA
	NCBI PROJECT NAME	Actinosynnema mirum DSM 43827
	GENUS	Actinosynnema
	GENUS SYNONYM	
	SPECIES	mirum
	SPECIES SYNONYM	
	STRAIN	101, DSM 43827
MIGS 13	CULTURE COLLECTION	DSM 43827 ATCC 29888 NBRC 14064
	TYPE STRAIN	Yes

GENOME PROJECT INFORMATION		
	GOLDSTAMP	Gc01024
	AVAILABILITY	Public
	SEQUENCING CENTERS	DSMZ DOE Joint Genome Institute
	FUNDING	DSMZ DOE
MIGS 38.1	CONTACT NAME	Bruce D
MIGS 38.2	CONTACT EMAIL	microbes@cuba.jgi-psf.org
	COMPLETION DATE	2009-06-09
	PUBLICATION	SIGS
	PROJECT RELEVANCE	Antibiotic production, Biotechnological, GEB

EXTERNAL LINKS		
	NCBI TAXON ID	446462
MIGS 1.1	NCBI PROJECT ID	19705
	GCAT_ID	002712_GCAT
	IMG OBJECT ID	643886170
	GENOME DATA	NZ_ABTA00000000
MIGS 33	GENOME DATABASE	NCBI IMG-GEBA DOE Joint Genome Institute

HOST METADATA		
ORGANISM METADATA		
MIGS 22	OXYGEN REQUIREMENT	Aerobe
MIGS 37.2	MOTILITY	Nonmotile
MIGS 37.3	SPORULATING	Sporulating
MIGS 37.12	TEMPERATURE RANGE	Mesophile
MIGS 15	BIOTIC REALTIONSHIPS	Free living
	DISEASE	None
	HABITAT	Soil
	TEMPERATURE	10-30
MIGS 37.11	PHENOTYPE	Nocardicin A producer
	ENERGY SOURCE	Chemoorganotroph

SEQUENCING INFORMATION		
MIGS 31.1	SEQUENCING STATUS	Complete
	SEQUENCING QUALITY	Level 6: Finished
MIGS 28.1	LIBRARY METHOD	8Kb, fosmid
MIGS 28.3	VECTOR	pMCL200, pcc1Fos
MIGS 30	ASSEMBLY METHOD	Newbler version 1.1.02.15, Phrap
MIGS 31.2	SEQUENCING DEPTH	8.9x Sanger; 20x pyrosequence
MIGS 34	GENE CALLING METHOD	Prodigal
MIGS 11	SIZE(kb)	8248
	ORFS	7177
MIGS 9	CHROMOSOME COUNT	1
MIGS 35	GC Content	73.7 %
MIGS 29	SEQUENCING METHOD	Sanger, Illumina, 454-FLX
	SEQUENCING COUNTRY	USA

ENVIRONMENTAL METADATA		
MIGS 4.5	ISOLATION SITE	"Blade of grass from Raritan River NJ, USA"
MIGS 5	COLLECTION DATE	September 1976
MIGS 4.6	ISOLATION COUNTRY	USA
MIGS 4	GEOGRAPHIC LOCATION	NJ USA
MIGS 4.1	LATITUDE	40.491816
MIGS 4.2	LONGITUDE	-74.322087
	PROJECT GEOGRAPHICAL MAP	

Figure 3. The GOLD CARD page. The GOLD CARD Page with the list of available metadata organized into six major categories. The corresponding MIGS/MIMS IDs are also shown for each GOLD field.

this flood, a new project management system (IMG-GOLD) was created to interface between GOLD and the Integrated Microbial Genomes (IMG) system (14). IMG is a widely used community resource for comparative analysis of publicly available genome data. The Expert Review version of IMG (IMG ER) (16) allows users to enter their own genome sequence data sets so that they can review and curate the annotations prior to their public release. Metadata accompanying those genome data sets are now captured via the IMG ER submission site (http://img.jgi.doe.gov/submit) and recorded in the new IMG-GOLD system (http://img.jgi.doe.gov/gold). IMG-GOLD now serves not only as the database underlying GOLD, but also as the source of metadata for IMG and IMG ER (and their metagenome counterparts, IMG/M and IMG/M ER). An example of how the metadata from GOLD can support and be presented through a metagenome analysis system, such as IMG/M (13), is presented in Supplementary Figure S1 (Supplementary Data). We anticipate that similar data exchange and interoperability between GOLD and other analytical systems, such as RAST (17) and CAMERA (18), will be developed in the near future.

Other systems already powered by GOLD include the NIH-funded Human Microbiome Project Catalog (HMPC) provided through the Data Analysis and

Coordination Center (DACC) (http://www.hmpdacc.org/). DACC connects directly to the GOLD database and accesses the HMP-specific data subset. To enable monitoring of the status of an HMP genome project, a new set of attributes and data types were added to GOLD and the already-existing controlled vocabularies were expanded. The HMPC page enables the DACC collaborators to choose and view targeted genome strains for sequencing. However, the community can also use this page to query the reference genomes and return profuse metadata.

IMG-GOLD also provides a web-based data entry mechanism that enables genome project submitters and curators to create/update/delete GOLD genome projects, provide associated metadata and create/edit controlled vocabularies for new metadata attributes. For users who prefer to provide metadata in file format, preformatted Excel spreadsheets are provided on the GOLD site (http://genomesonline.org/Project_submission.htm) for both genome and metagenome projects.

Data availability

All GOLD data are available according to the Creative Commons License of Attribution-NonCommercial-Share Alike (http://creativecommons.org/licenses/by-nc-sa/2.5/). All of the available metadata types in GOLD can be downloaded to an Excel file to facilitate wider distribution and use of the data.

OVERVIEW STATISTICS

Several different types of statistics, related to each of the data fields, can be derived using GOLD's advanced search engine, the new metadata search capability, and the data download capability. In addition, graphical overviews for specific data types are provided via the 'Gold Statistics' link on the database home page (http://genomesonline.org/gold_statistics.htm). This feature is supported for the data fields discussed in the following paragraphs.

Evolution of genome projects

Genome project tracking in GOLD has been steadily increasing over time with an average\2.25-fold increase every 2 years for the past 12 years (Figure 1A). The microbial genome projects have been carrying the majority of that increase. This systematic and comprehensive genome project tracking can help addressing two major questions: (i) where and how numerous are the remaining gaps in sequencing along the bacterial and archaeal branches of the tree of life, and (ii) how accurately can we predict the number of genome projects that will be sequenced over the next 3–5 years?

Table 4 addresses the first question by reporting the taxonomic distribution of genome projects, showing for each taxon the number of genome projects compared with the total number of described taxonomic units (filtering out the environmental and the unknown entries). In effect, it identifies the taxonomic groups in each domain of life for which there are no currently registered genome projects. These taxonomic groups should eventually become targets for new sequencing projects. Further, we hope that the availability of this systematic project monitoring will not only help identify the next sequencing targets, but also help the sequencing centers to avoid unnecessary redundancy and duplication of efforts.

Table 4. Taxonomic distribution of genome projects[a]

Domain	Phyla	Class	Order	Family	Genus
Archaea: **179**	**5**/5	**9**/9	**24**/26	**24**/26	**85**/109
Bacteria: **4184**	**27**/29	**45**/47	**234**/281	**234**/281	**730**/1930
Eukarya: **1280**	**29**/55	**80**/188	**350**/6288	**350**/6288	**536**/47906

[a]Available at: http://genomesonline.org/phylogenetic_distribution.cgi. For each taxon, the number with genome projects (bold) compared to the total number of identified taxons according to NCBI's Taxonomy.

Table 5. Predicted increase of microbial genome sequencing projects

	1995–2009	2010–2015
Finished	1000	3000
Draft	1100	11 000
Genes	7.5 million genes	56 million genes

Table 5 attempts to address the second question which is what is the anticipated growth of the microbial genome projects over the next 5 years? Following a very conservative estimate we would expect to see three times increase in the number of the complete and 10 times in the number of the draft microbial genome projects that have been sequenced during the last 15 years. However, if we extrapolate a linear increase in the number of finished and draft genomes based on Figure 1A, those predictions would be realized within the next 3 years.

Sequencing centers

Four major sequencing centers account for about 50% of the 5843 sequencing projects currently monitored in GOLD (Figure 1B), a situation that has not changed over the last 2 years. However, when considering only archaeal and bacterial projects, the two leading sequencing centers (JGI and JCVI) now represent a smaller share: about 35%, compared to more than half 2 years ago. The fact that a much larger community is now carrying out these projects compared to 2 years ago also reflects the increasing democratization of the sequencing technology.

Phylogenetic distribution

The sampling bias favoring three major bacterial lineages —Proteobacteria, Firmicutes and Actinobacteria—has decreased only slightly during the last couple of years (Figure 1D). The above three lineages now comprise 80% of all genome projects compared to 82% 2 years ago. This small shift is due mostly to large-scale sequencing efforts, such as the GEBA and HMP, which target previously neglected phylogenetic lineages. Clearly, there remains much room for improvement here, and further progress can

be expected if similar large-scale biodiversity sequencing efforts continue.

FUTURE DIRECTIONS

The challenges facing GOLD have increased dramatically as GOLD continues to evolve from a genome/metagenome project monitoring system into a universal genome project core catalog/indexer charged with the task of providing data interconnectivity, exchange and dissemination. In this new role, GOLD is required to efficiently store, process and automatically track metadata that is rapidly increasing in scope and complexity. All the while, there is a great expectation for GOLD to pioneer future genomic standards.

To meet these challenges will require the creation of a shared genome project conceptual model and a database schema to handle the genome-project-associated metadata. The genome/metagenome data continue to be somewhat structured and hierarchical, but the rich associated metadata information becoming available requires the creation of a 'Genome Project Ontology' for effective management. Incorporation of other available ontologies, such as existing medical and environmental ontologies, is part of the immediate plan.

Furthermore, numerous other bioinformatics databases and researchers will need to acquire and/or synchronize with GOLD data. To address their needs, GOLD will provide access for client programs via web services using SOAP, GOLDXML and other RESTful technologies, as well as communicate with subscribers via RSS feeds. To further increase community access to GOLD, a GOLD-wiki site will be established where genome project curators can contribute additional project information using various media-rich data formats. We also plan to employ data warehousing tools to facilitate reporting and analysis of the GOLD data on the statistics page, thereby eliminating the need for the manual creation of Excel charts that become quickly outdated. To improve data mining, the GOLD search engine will provide an advanced query mechanism wherein the search criteria available will depend on the meta-properties of the input objects.

DATABASE AVAILABILITY

GOLD can be accessed at: http://www.genomesonline.org/.

Further comments and feedback are welcome at: mail@genomesonline.org.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Merry Youle for her excellent editorial assistance. GOLD has been maintained and developed mostly based on the volunteer work of its small team. We are grateful to all the colleagues who kindly provide information for the more accurate monitoring of the genome projects and particularly to Michelle Giglio and Heather Huot from University of Maryland. The support of Rashida Lathan, Stella Proukaki and Tatiana Drakakis is especially acknowledged. The full list of all contributors is available at: (http://www.genomesonline.org/acknowledgments.html).

FUNDING

The US Department of Energy's Office of Science, Biological and Environmental Research Program; and by the University of California, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344; and Los Alamos National Laboratory under Contract No. DE-AC02-06NA25396. Funding for open access charge: Department of Energy.

Conflict of interest statement. None declared.

REFERENCES

1. Kyrpides,N. (1999) Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*, **15**, 773–774.
2. Bernal,A., Ear,U. and Kyrpides,N. (2001) Genomes Online Database (GOLD): A Monitor pf genome projects world-wide. *Nucleic Acid Res*., **29**, 126–127.
3. Liolios,K., Tavernarakis,N., Hugenholtz,P. and Kyrpides,N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of Genome Projects world-wide. *Nucleic Acid Res*., **34**, D332–D334.
4. Liolios,K., Mavromatis,K., Tavernarakis,N. and Kyrpides,N.C. (2007) The Genomes OnLine Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acid Res*., **36**, D475–D479.
5. Field,D., Garrity,G., Gray,T., Morrison,N., Selengut,J., Sterk,P., Tatusova,T., Thompson,N., Allen,M.J., Ashburner,M. *et al*. (2008) Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification. *Nat. Biotechnol*., **26**, 541–547.
6. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2007) GenBank. *Nucleic Acids Res*, **35**, D21–D25.
7. Kulikova,T., Akhtar,R., Aldebert,P., Althorpe,N., Andersson,M., Baldwin,A., Bates,K., Bhattacharyya,S., Bower,L., Browne,P. *et al*. (2007) EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res*., **35**, D16–D20.
8. Okubo,K., Sugawara,H., Gojobori,T. and Tateno,Y. (2006) DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res*., **34**, D6–D9.
9. Birney,E., Hudson,T.J., Green,E.D., Gunter,C., Eddy,S., Rogers,J., Harris,J.R., Ehrlich,S.D., Apweiler,R., Austin,C.P. *et al*. (2009) Prebublication data sharing. *Nature*, **461**, 168–170.
10. Kyrpides,N.C. (2009) Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nat. Biotechnol*., **27**, 627–632.
11. Garrity,G.M., Field,D. and Kyrpides,N.C. (2009) Standards in genomic sciences. *Stand. Genomic Sci*., **1**, 1–2.
12. Dawyndt,P., Vancanneyt,M., DeMeyer,H. and Swings,J. (2005) Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 1111–1126.
13. Markowitz,V.M., Ivanova,N.N., Szeto,E., Palaniappan,K., Chu,K., Dalevi,D., Chen,I.-M.A., Grechkin,Y., Dubchak,I., Anderson,I. *et al*. (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res*., **36**, D534–D538.

14. Markowitz,V.M., Szeto,E., Palaniappan,K., Grechkin,Y., Chu,K., Chen,I.-M.A., Dubchak,I., Anderson,I., Lykidis,A., Mavromatis,K. *et al.* (2008) The Integrated Microbial Genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res.*, **36**, D528–D533.
15. Hirschman,L., Clark,C., Cohen,K.B., Mardis,S., Luciano,J., Kottmann,R., Cole,J., Markowitz,V., Kyrpides,N., Morrison,N. *et al.* (2008) Habitat-Lite: a GSC case study based on free text terms for environmental metadata. *OMICS*, **12**, 129–136.
16. Markowitz,V.M., Mavromatis,K., Ivanova,N.N., Chen,I.-M.A., Chu,K. and Kyrpides,N.C. (2009) Expert IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics*, **25**, 2271–2278.
17. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E.M., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. *et al.* (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **19**, 386.
18. Seshadri,R., Kravitz,S.A., Smarr,L., Gilna,P. and Frazier,M. (2007) CAMERA: a community resource for metagenomics. *PLoS Biol.*, **5**, e75.